International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.9, No.12, (2016), pp.123-134
http://dx.doi.org/10.14257/ijsip.2016.9.12.13

Algorithmic Research for Image Annotation Based on Region Convolution Neural Network

Yuan Yuli

College of Computer Science, Neijiang Normal University
yuanyuli@yeah.net

Abstract

Traditional image processing and pattern recognition research aimed at identifying the target image. As time goes by, on the basis of the recognition of the image, more and more research points to identify multiple targets in the image, and the corresponding block of the corresponding target is calibrated. Compared with the traditional image recognition, the problem of image annotation is a combination of multi classification and multi regression, which is more difficult and challenging. On the basis of deep convolutional neural network, this paper studies the image annotation algorithm based on region selection algorithm and support vector machine, and the algorithm is tested on the PASCAL VOC 2010 image data set. Experimental results show that compared with the current algorithm, this algorithm can be used to mark the image of multiple targets, the effect is obvious, and there is a great practical significance.

Keywords: *Image annotation; Outline box; Convolutional neural network; Support vector machine*

1. Introduction

In the field of image processing and pattern recognition, the research effort has transformed from image recognition to image annotation step by step [1]. According to machine learning and pattern recognition algorithm, the traditional image recognition can finish the work of feature extraction and classification. At last, the label of each input image is given to complete the recognition of the image. On this basis, under normal circumstances, the target in the scene image will appear repeatedly, and using a label to identify an image cannot complete the recognition of the scene image. In order to better identify multiple targets in the image, people began to identify the location of each target and its tags, and then use the methods of image processing and pattern recognition to solve the problem of image annotation [2].

Traditional image recognition consists of two parts: feature extraction and classification recognition. In the aspect of feature extraction, the researchers used the basic characteristics of the image, including color feature, texture feature and shape feature [3]. With the further research, more and more advanced features are applied to image recognition, including SIFT features, SURF features, multi scale sparse features and so on [4]; in the aspect of classification recognition, the early machine learning models have achieved good results, including K nearest neighbor, clustering, support vector machine, artificial neural network, naive Bayesian method and so on [5]. In recent years, with the deepening of research, more and more researchers use in-depth learning model for feature extraction and recognition, which has made great progress in image recognition.

Similar to traditional image recognition, image annotation also requires two basic steps: feature extraction and classification regression. Early image annotation methods use the basic image features, image classification and regression method mentioned in the image recognition. In recent years, the depth learning model can automatically complete the

ISSN: 2005-4254 IJSIP

feature extraction, it is a kind of feature extraction algorithm, which can obtain a more stable recognition results. At present, depth of learning models made a lot of progress only in the image recognition. The difference is that the image annotation problem cannot only input image in the input, but also need to enter the target part of the image as a training or test samples, which brings a lot of difficulty to the deep learning framework. Based on this, under the framework of deep learning model, we first get the annotation candidate box of the image rely on the selective search algorithm, and then put it into the neural network to extract features. Finally, the extracted features and corresponding labels are input into the support vector machine for regression, the regression results are given.

2. Algorithm Framework of Image Annotation Based on Region Convolution Neural Network

In 2006, depth learning model [6] was proposed by Hinton Geoffrey and his students through analyzing the shortcomings of traditional neural network. In fact, deep learning uses the hierarchical structure which is similar to the traditional neural network, and a deep learning system consists of three layers: input layer, hidden layer and output layer.

In this paper, depth learning model is used to classify and return multiple targets. The depth model is a branch of PRML, which is developed from neural network. The sign of depth model is that the model has multiple hidden layers as middle layer, it extracts a wide variety of image features which have the nature of descriptive meaning through the hierarchical relationship between multiple hidden layers, and it is the greatest advantage of the depth learning model. Depth learning model is no longer needed to design and extract complex image features; the whole model will automatically extract a reasonable feature for image classification according to the parameters. In addition, multiple hidden layers of neural networks are more consistent with the structure of human brain neurons; such a system structure is more suitable for classifying and identifying complex things.

Based on convolutional neural network, we first get the annotation candidate box of the image rely on the edge box algorithm, and then put it into the neural network to extract features. Finally, the extracted features and corresponding labels are input into the support vector machine for regression, the regression results are given. The following Figure 1 gives the algorithm of this paper-- framework of image annotation algorithm based on region convolution neural network.

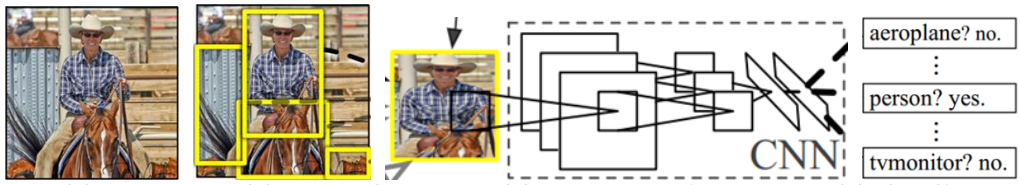


Figure 1. Framework of Image Annotation Based on Region Convolution Neural Network

3. Algorithm Elaboration of Image Annotation Based on Regional Convolutional Neural Network

3.1. The Acquisition of Image Annotation Candidate Frame

In the image annotation candidate selection, this paper uses the “Edge boxes” [7] to obtain the candidate region. In this algorithm, we first extract the contour information of the original image, and give a number of annotation boxes, then calculate the number of internal contours in each annotation box and the number of contours that intersect with the edge of the border. When the numbers of contours in a label box outline much more

than the numbers of contours at the edge of the frame, the annotation box very likely contains a target. All the labeled boxes are sorted and classified according to this information,, and the scores of each mark frame are obtained. Then the target is selected as the candidate frame for image annotation.

The algorithm consists of four parts:

(1) Input the original RBG image, filter the noise and other pretreatment operation;

(2) Contour extraction, since the goal is the larger contour, some small outline is removed by using sparse algorithm to obtain a stable large profile.

(3) Traverse all the contours, the contour points in the contour which almost in the same line are gathered to form a set of contours and different colors are used to represent them. After the contour set is formed by the same line contour, the N contour set can be acquired, and then the similarity between each contour set can be calculated by the following Formula (3-1).

$$a(s_i, s_j) = |\cos(\theta_i - \theta_{ij}) \cos(\theta_j - \theta_{ij})|^{\gamma} \quad (3\text{-}1)$$

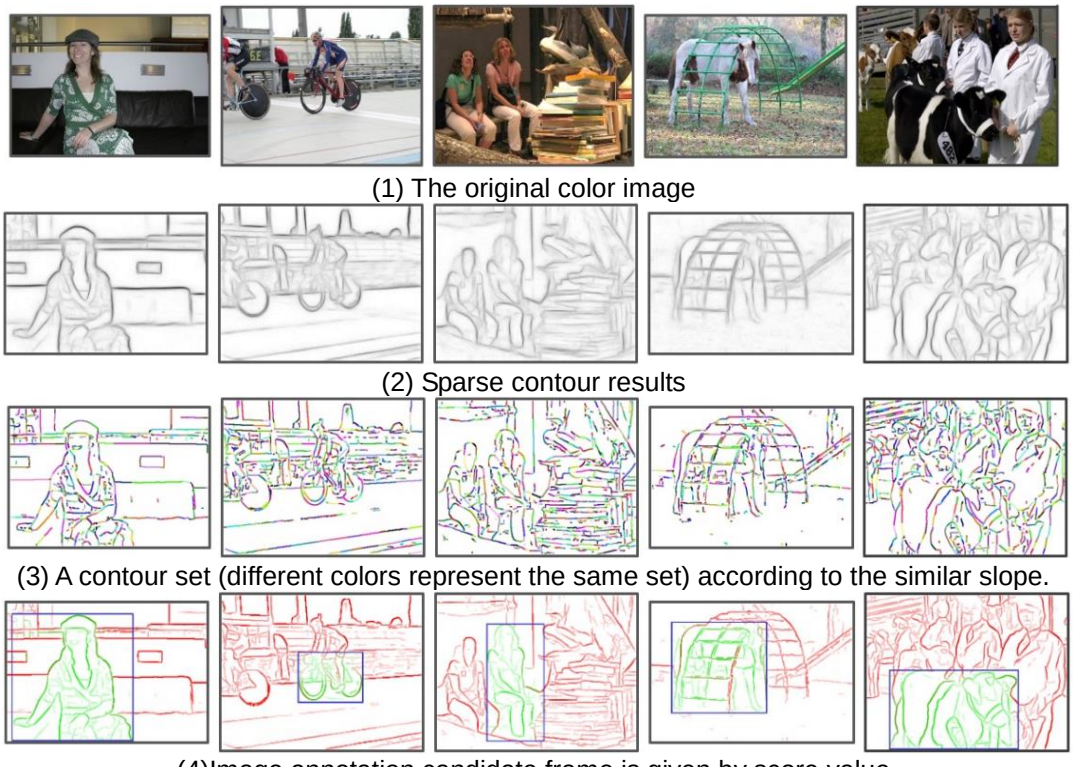
Among them, θ_{ij} represents the included angle between i and j, which is used to enhance the connectivity between i and j.

(4) In the end, use the (3-2) to calculate whether the score value of the contour is in the contour set and keep the score high as the correct image annotation candidate box.

$$w_b(s_i) = 1 - \max_T \prod_{j=0}^{|T|-1} a(t_j, t_{j+1}) \quad (3\text{-}2)$$

T represents the set sequence value from the edge of the candidate frame to S_i.

The following Figure 2 is the four step of the algorithm that extract the image annotation candidate box.



(1) The original color image

(2) Sparse contour results

(3) A contour set (different colors represent the same set) according to the similar slope.

(4)Image annotation candidate frame is given by score value

Figure 2. Flow and Results of Contour Box Algorithm

3.2. Feature Extraction of Candidate Frame for Image Annotation

The training set and test set of PASCAL VOC 2010 data set are extracted by using the outline box algorithm. Then, each of the labeled boxes is cut out and input into the convolutional neural network for feature extraction.

3.2.1. Framework of Convolutional Neural Network: Convolutional neural network [8] developed from the neural network. The traditional neural network includes input layer, hidden layer and output layer. The convolutional neural network adds multiple hidden layers to improve the quality of the extracted image features. In this paper, we use the neural network architecture, which consists of 5 convolution hidden layer and 3 fully connected layer. Input three channel image 224*224*3 dimensional data, neural network extract 4096 dimensional feature as the output, and the results of the output is also the results that extracted by convolution neural network. Figure 3 is the architecture diagram of convolutional neural network.

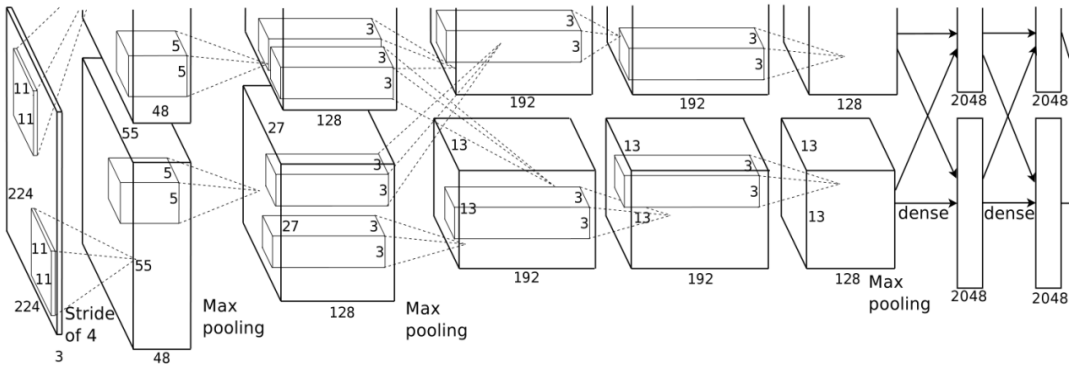


Figure 3. Architecture Diagram of Convolutional Neural Network

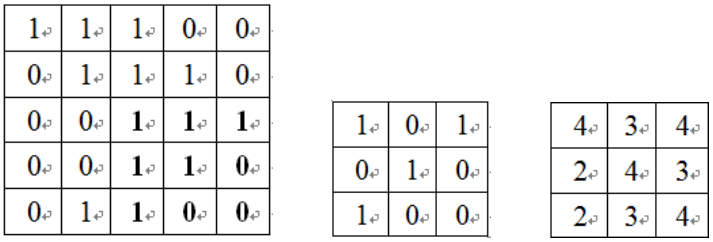
The graph is composed of input layer, 5 layers, 3 full volume link layers and output layer. The input image is 224*224*3 dimension data with the uniform size, and the output is 4096 dimensional data feature. Pool operations are used after the first, the second and the fifth convolution layer, which reduce the dimension of the convolution results and the scale parameter. Because of too many input image, this paper use parallel convolution kernel pool operation simultaneously in two GPU, and exchange data between the second and third volumes layers which further improve the time complexity of training image annotation.

3.2.2. Convolution Operation: Convolution operation is one of the most important operations in the convolutional neural network, which can be used to extract the features before the new combination of convolution kernel rules to obtain more complex features. As is known to all, the complexity of an image can be combined by using simple features in some way. In the convolutional neural network, the convolution operation can transform a single image feature into a more complex image feature, and the more complex the image features can be used for classification. Digital image is stored through the two-dimensional discrete data. Assuming the original image is f (x, y), the convolution kernel is g (x, y), and then the convolution can be defined as:

$$f(x,y)*g(x,y)=\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f(m,n)g(x-m,y-n)x=0,1,2...M-1,y=0,1,2...,N-1 \qquad (3\text{-}3)$$

The convolution computation of two-dimensional images can be mapped to the corresponding convolution value of the continuous sliding convolution window. Figure 4

shows the convolution operation of the image with the convolution kernel.



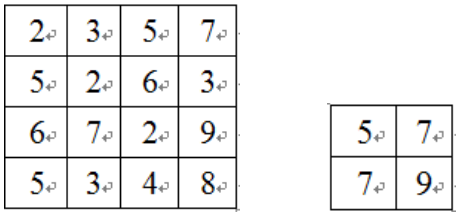
(a) Image (b) Kernel (c) Convolution

Figure 4. Convolution Operation Diagram

The convolution kernel in the image above is g=[1,0,1; 0,1,0; 1,0,1]. In the practical training of convolutional neural networks, each layer has a number of convolution kernel, the purpose is to get more extensive combination of features to improve the recognition rate of image features. For example, in this paper, the convolution operation in the first layer use 256 different convolution kernels and the second convolution layer will produce 256 different convolution results for subsequent using.

3.2.3. Pool Operation: The sharing of convolution kernel can effectively reduce the training parameters, but with the increase of the number of the convolutional layers, the corresponding neural network parameters will increase exponentially. For example, under the five layers of convolutional layer, it will produce about 100 million parameters, many redundant parameters not only increases the training time but also cause the over fitting [9], which lead to a decline in the rate of image recognition. So this paper adds the pool layer after the convolution layer to reduce the parameters of neural network.

The pool operation is used to reduce the parameters in all regions by calculating the statistical features of a region and consider it as the representative of the region's features. Because the static image generally has the aggregate attribute, the characteristic in some area also can have certain effect in other regions. So for more neural network parameters, we can aggregate multiple parameters to get fewer parameters, but the parameters are able to represent the specific characteristics of the region. The common methods are mean pooling and maximum pool. Mean pooling takes the mean of selected area as the characteristic of the region; maximum pool takes the maximum of selected area as the characteristic of the region. Similar to convolution, pooling is also performed on the entire image by a sliding window, as shown in Figure 5, which gives the pool operations.



(a) Image (b) Pooling

Figure 5. Pool Operation

Pool operations get the results of 1*1 by an averaging process, which reduces the number of parameters. The author uses mean pooling to add the pool layer after the convolution results of the first layer, the second layer and the fifth layer, and then reduce the training parameters from 1 million to 600 thousand. Through the experimental results, we can see that the situation of over fit is less after pooling and it is more suitable for image classification and recognition.

3.3.4. Update Weight Based on Back-Propagation: Like traditional neural network, convolutional neural network update the parameters of neurons in each layer by back-propagation, optimize various weights of neural network by constant iteration, so that the predicted value of the neural network can be close to the actual value of the tag. Back propagation network is one of the most commonly used optimization methods to minimize the error. It is a strategy of updating the parameters by solving the square error of the forward output tag value and the actual label value. Assuming that there are a total of C categories, N samples, where t_k^n represents the actual tag value, y_k^n on behalf of the tag value of the neural network. Then we can use the formula to show the square error cost function:

$$E = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{c}(t_k^n - y_k^n)^2 = \frac{1}{2}\left\|t^n - y^n\right\|^2 \tag{3-4}$$

According to the forward propagation of the neural network we can know that in order to update the weights, the weight of ω^l and bias b^l in the first layers of neural networks can be calculated by the following formula.

$$x^l = f(u^l), u^l = \omega^l x^{l-1} + b^l \tag{3-5}$$

Solving the parameters ω^l and b^l by the variation of the error E, the stochastic gradient descent method is used to solve the partial derivative of b^l.

$$\frac{\partial E}{\partial b} = \frac{\partial E}{\partial u}\cdot\frac{\partial u}{\partial b} = \delta \tag{3-6}$$

$\frac{\partial u}{\partial b} = 1, \frac{\partial E}{\partial b} = \frac{\partial E}{\partial u}$, then reverse the spread $\frac{\partial E}{\partial b}$ continuously according to the formula (3-5) , it can spread the error to the bottom, and the error rate of change δ^l can be derived from the first layer according to (3-3) and (3-4):

$$\delta^l = (\omega^{l+1})^T \delta^{l+1} \circ f'(u^l) \tag{3-7}$$

The error in the l+1 layer spreads to the L layer spread, and eventually spread to the first volume layer. Combined (3-5), (3-6) and (3-7) we can obtain the derivative between the error E and the weight of the ω^l.

$$\frac{\partial E}{\partial \omega^l} = x^{l-1}(\delta^l)^T \tag{3-8}$$

In this way, the weights ω^l and offset b^l in each layer can be iteratively updated by the partial derivative $\frac{\partial E}{\partial \omega^l}$ and $\frac{\partial E}{\partial b^l}$ of error E, until the minimum error of the weights is obtained, and the training of the convolutional neural network is completed.

3.3. Regression Result of Image Annotation

The convolutional neural network is transformed into the feature vector of the 4096 dimensional feature extracted from each candidate region, and the training set and test set of PASCAL VOC 2010 data set are extracted. Then, the extracted feature vectors and the corresponding label are input into support vector machines to complete the final label result regression.

3.3.1. Classified Regression of Support Vector Machine: Support vector machine

(SVM) [10] developed from the perceptron, traditional perception machine [9] separate the data by solving linear classification. The data distribution is different, and the linear classification is more, but the support vector machine needs to find an optimal classification surface, so that the data to be classified can be separated from the maximum "interval". Traditional support vector machines are mainly aimed at the binary classification. A large number of studies show that multi category classification problem can be simplified into multiple binary classification problems. In order to solve the image annotation problem, we only need to create multiple support vector machine classification.

Assume that the training set $\{x_i, y_i, \mathrm{i} = 1,2,3\ldots \mathrm{n}\}$ represents a data set, Data x is the sample data needed to classify, while y is corresponding label of each data. It is also a description that which category the data needs to be classified. Support vector machines need to find an optimal classification face w*x+b=0, so that the sample data set x can be separated as large as possible. As shown in Figure 6, we need to separate the two categories of data as far as possible, which is the best classification shown in the graph.

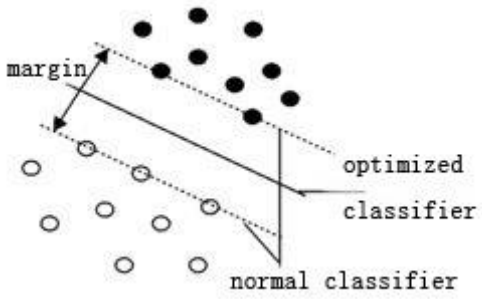


Figure 6. Support Vector Machine and Optimal Classification Face

By analyzing the optimal classification surface w*x+b=0, where w is the weight, x is the support vector, and b is the additive bias. Calculating the classification face of optimization is also the calculation of two plans for w and b:

$$\min_{\omega,b} \frac{1}{2}\|\omega\|^2 \qquad (3\text{-}9)$$

$$s.t. y_i(\omega \square x_i + b) \geq 1, i = 1,2,...,n \qquad (3\text{-}10)$$

By solving the optimization problem of the two programming, the optimal weights w and b can be solved by the Lagrange multiplier method and the KKT condition, and then the optimization results of the sample classification are obtained.

3.3.2. The Effect of Support Vector Machine Enhanced by Kernel Method: Actually, almost all practical problems cannot meet linearly separable data set x just like Figure 2-1. Linearly separable data set cannot just use the linear classification to distinguish them. On the basis of the support vector machine, adding the slack variables, and allowing partial sample emerge errors, in this case we can use the support vector machine method for solving nonlinear separable data sets. In addition, as shown in Figure 7, such a data set cannot be solved through the slack variables. Because if it is classified in accordance with the linear classification plane, then there will be more wrong samples, which leads to the poor performance of the final classifier. At this time, the support vector machine uses the concept of kernel function [11] .The original linear non separable data sets can be changed into the high dimension space through the kernel function, and the original data is transformed into a linear separable data set in high dimensional space through high

dimensional space mapping. In this way, we can use the linear classification to classify.

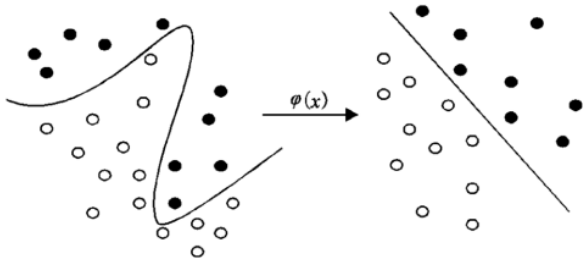


Figure 7. Nonlinear Data Sets Can be Divided into High Dimensional Linear Separable Data Sets by Kernel Function

Among them, φ_x is the kernel function and we can transform the two optimization problem of support vector machine to the following:

$$\min_{\omega,b} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m}\xi_i \quad (3\text{-}11)$$

$$s.t.y_i(\omega \cdot \varphi(x_i)+b) \geq 1-\xi_i \quad (3\text{-}12)$$

$$\xi_i \geq 0, i=1,2,3,...,n \quad (3\text{-}13)$$

ξ_i is a relaxation factor, φ_x is the kernel function, and C is the penalty factor. In the same, we can use the Lagrange multiplier method and the KKT condition to get the optimal solution of the two programming, and then show the classification surface of the nonlinear support vector machine:

$$f(x) = \mathrm{sgn}[\sum_{i,j=1}^{m} a_i y_i K(x_i, x_j) + b] \quad (3\text{-}14)$$

$K(x_i, x_j)$ is a kernel function, we can easily use the support vector machine with kernel function to complete the regression problem of image annotation frame by using kernel function.

4. Experiment and Result Analysis

In order to verify the feasibility and effectiveness of the region convolution neural network (Region-CNN) algorithm in this paper, we test the algorithm on image annotation data set PASCAL VOC 2010 [12]. The results of the IOU value are generally used to reflect the results of the image annotation. The value interval is [0, 1], which is a value that describes the results of the annotation box. When the overlapping between result frame and target is larger, the results of IOU are smaller. On the contrary, when the description frame is able to surround the target, the larger the result of IOU, the better the effect is. The results of regional convolutional neural network are described for image annotation. In general, when the value of IOU is greater than 0.7, the corresponding annotation box can be a good reflection of the results. So, the following Figure 8 gives the results that IOU is more than 0.7.

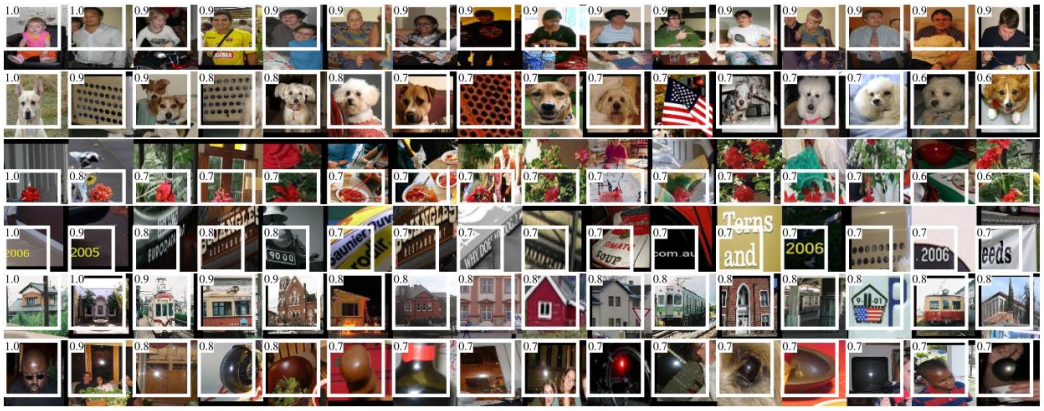


Figure 8. Image Annotation Results (IOU>0.7)

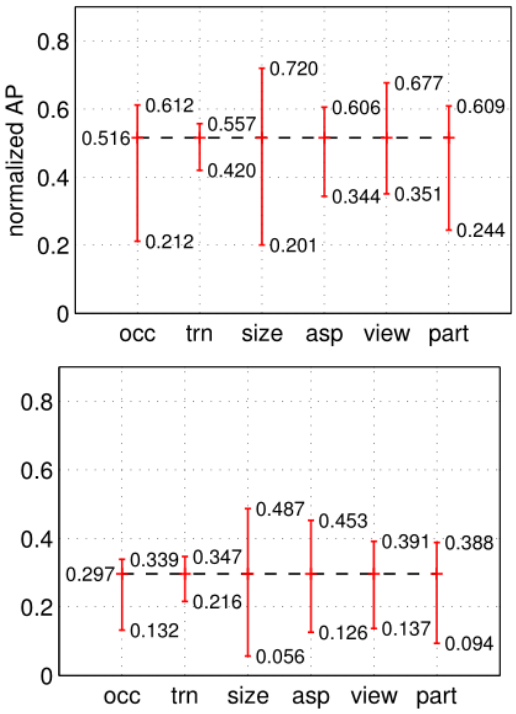
From the annotation results can be seen, when IOU>0.7, corresponding annotation box can well express the labeling of the target. That is, the support vector machine algorithm in this paper can return label box to the marking target.

For horizontal comparison, this paper compares with algorithm DPM [13], UVA algorithm [14], Regionlets algorithm [15] and SegDPM algorithm [16], and uses the same training data set and test data set to compared, the IOU of the results are set 0.7. The following table 1 gives the horizontal comparison between this algorithm and the other algorithms. Test results show the correct values of the target results. Include (dog, cat, bird, ball, sheep, people, tree, mirror, bus, car, horse). The average correct rate (mAP) results of several common annotation results are given.

Table 1. Comparison of Experimental Results between this Algorithm and Several Traditional Algorithms

VOC 2010	dog	cat	bird	ball	sheep	people	tree	mirror	bus	car	horse	mAP
DPM	49.7	55.6	42.9	54.7	44.6	29.8	65.8	10.9	14.8	49.9	38.6	41.57
UVA	53.2	57.4	41.9	49.7	54.7	32.6	68.7	13.8	16.7	52.7	39.1	43.68
Regionlets	55.6	54.9	44.8	55.4	43.5	39.7	66.9	17.8	22.4	55.6	40.7	45.21
segDPM	56.7	55.4	46.8	58.7	50.4	44.8	67.9	18.6	22.7	54.2	43.8	47.27
Region-CNN	64.8	58.9	49.8	58.4	55.6	49.7	75.4	23.1	30.4	60.8	55.5	52.95

According to the algorithm results in the PASCAL VOC 2010 data sets can be seen that, for traditional methods including variable part of the model, algorithm in this paper can exceed its results to the common target recognition, and average correct rate can get significant improvement. Compared with the average correct rate of traditional DPM algorithm, lifting rate exceeds 50% for the most common number of targets. Finally, the algorithm is able to obtain a significant average accuracy of 53.6% on the test set of the PASCAL VOC 2010 data set. However, compared with the traditional methods, this paper puts forward the depth study method, the weights number of the convolutional neural network is more, the consumption of training time is larger while the training resources consumption of the traditional methods is relatively low. However, in the process of testing, it only costs the time of candidate frame extraction, feature extraction and regression.



(a)Our Improved algorithm (b)Conventional DPM algorithm

Figure 9. Comparison Chart of Average Correct Rate of the Algorithm and the Traditional DPM Algorithm

5. Conclusions

With the development of computer vision, image processing, pattern recognition and artificial intelligence, more and more scholars have devoted themselves to the research of image processing. Different from the traditional image recognition, this paper aims to study the image annotation algorithm, identify different objects from the image, and give the corresponding annotation frame. In order to better identify the object, this paper uses the depth study of convolutional neural network model. First of all, the author uses edge information from the original image to extract annotation objects in the candidate region, then subsequent target regional are input to the convolutional neural network to feature extraction of fixed length. Finally, the features of fixed length and tag are input to support vector machine for regression. Regression results are given to show the results of image annotation. The result of the image annotation is 53.6% on the PASCAL VOC 2010 data set. Compared with the traditional DPM method, the result has achieved good results, and has a strong practical and research value.

Reference

[1] J. Johnson, L. Ballan and L. Fei-Fei, "Love thy neighbors: Image annotation by exploiting image metadata", Proceedings of the IEEE International Conference on Computer Vision, **(2015)**, pp. 4624-4632.
[2] G. Lev, G. Sadeh and B. Klein, "RNN Fisher Vectors for Action Recognition and Image Annotation", arXiv preprint arXiv:1512.03958, **(2015)**.
[3] R. V. Sharan and T. J. Moir, "Cochleagram image feature for improved robustness in sound recognition", Digital Signal Processing (DSP), 2015 IEEE International Conference on. IEEE, **(2015)**, pp. 441-444.
[4] A. P. Tafti, A. Baghaie and A. B. Kirkpatrick, "A Comparative study on the application of SIFT, SURF, BRIEF and ORB for 3D surface reconstruction of electron microscopy images[J]. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, **(2016)**, pp. 1-14.

[5] L. C. A. M. Cavalcanti, J. R. H. Carvalho and E. M.dos Santos, "A Comparison on Supervised Machine Learning Classification Techniques for Semantic Segmentation of Aerial Images of Rain Forest Regions", **(2015)**.
[6] J. Schmidhuber, "Deep learning in neural networks: An overview", Neural Networks, **(2015)**, vol. 61, pp. 85-117.
[7] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges", In ECCV, **(2014)**. pp. 1, 2, 3,4, 6, 7, 8
[8] M. Oquab, L. Bottou and I. Laptev, "Is object localization for free?-weakly-supervised learning with convolutional neural networks", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, **(2015)**, pp. 685-694.
[9] V. N. Murthy, S. Maji and R. Manmatha, "Automatic Image Annotation using Deep Learning Representations", Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. ACM, **(2015)**, pp. 603-606.
[10] H. Xiong, S. Szedmak and J. Piater, "Scalable, accurate image annotation with joint SVMs and output kernels", Neurocomputing, vol. 169, **(2015)**, pp. 205-214.
[11] W. Liu, H. Liu and D. Tao, "Manifold regularized kernel logistic regression for web image annotation", Neurocomputing, vol. 172, **(2016)**, pp. 3-8.
[12] C. Szegedy, S. Reed, D. Erhan and D. Anguelov, "Scalable, high-quality object detection", arXiv preprintnarXiv:1412.1441, 2014. 2
[13] R. Girshick, P. Felzenszwalb and D. McAllester, "Discriminatively trained deformable part models", release 5.
[14] J. Uijlings, K. van de Sande, T. Gevers and A. Smeulders, "Selective search for object recognition", IJCV, **(2013)**.
[15] X. Wang, M. Yang, S. Zhu and Y. Lin, "Regionlets for generic object detection", In ICCV, **(2013)**.
[16] S. Fidler, R. Mottaghi, A. Yuille and R. Urtasun, "Bottom-up seg- mentation for top-down detection", In CVPR, **(2013)**.

Authors

Yuan-Yuli, The author is a lecturer in the university; the research direction is the computer application and image processing. Work unit: school of computer science, Neijiang Normal University Address: School of computer science, Neijiang Normal University.

